(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2009-146406
(P2009-146406A)
(43)公開日　平成21年7月2日(2009.7.2)

(51)Int.Cl.		F I			テーマコード（参考）
G06T 7/20	(2006.01)	G06T	7/20	B	2F065
G01B 11/00	(2006.01)	G01B	11/00	H	5L096

審査請求 有　請求項の数 12　ＯＬ　外国語出願　（全 43 頁）

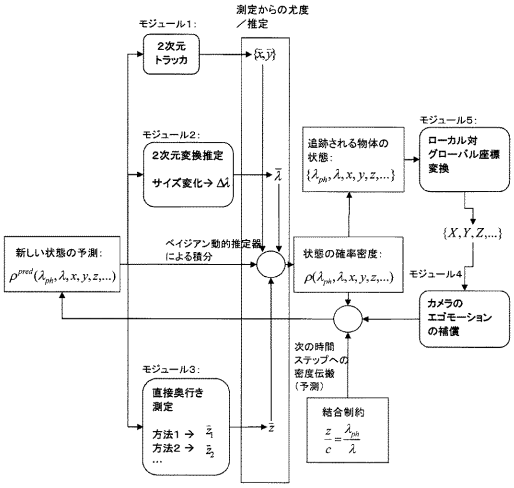
(21)出願番号　特願2008-314454 (P2008-314454)
(22)出願日　平成20年12月10日 (2008.12.10)
(31)優先権主張番号　07122829.0
(32)優先日　平成19年12月11日 (2007.12.11)
(33)優先権主張国　欧州特許庁 (EP)

(71)出願人　503113186
ホンダ　リサーチ　インスティテュート　ヨーロッパ　ゲーエムベーハー
Ｈｏｎｄａ　Ｒｅｓｅａｒｃｈ　Ｉｎｓｔｉｔｕｔｅ　Ｅｕｒｏｐｅ　ＧｍｂＨ
ドイツ連邦共和国　デー－６３０７３　オッフェンバッハアムマイン　カール・レギエン・シュトラーセ　３０
(74)代理人　110000246
特許業務法人オカダ・フシミ・ヒラノ
(72)発明者　ジュリアン・エガート
ドイツ国６３１７９　オーベルシャウゼン、ペテル－デルプ－シュトラーセ　３３

最終頁に続く

(54)【発明の名称】２次元見え方およびマルチキュー奥行き推定を使用する実世界の物体の視覚的追跡方法

(57)【要約】　（修正有）
【課題】カメラシステム、２次元画像情報、およびカメラから物体までの距離の様々な測定値の組合せを用いて、経時的な実世界物体の動的状態を推定するための方法を提供する。
【解決手段】２次元画像情報は、物体の見え方を用いて、物体の２次元位置のみならず、その２次元サイズおよび２次元サイズ変化をも追跡するために使用される。加えて、カメラから物体までの距離は、１つまたは幾つかの直接奥行き測定から得られる。２次元位置およびサイズ、ならびに物体の奥行きを相互に結合することにより、物体の３次元位置および３次元速度の改善された推定が得られ、したがって、動的視覚的場面解析用のカメラが搭載されたロボットまたは自動車のような移動するプラットフォーム上で使用することのできる、改善された実世界物体追跡システムが得られる。
【選択図】図１



【特許請求の範囲】
【請求項１】
２次元見え方ヒントおよびマルチキュー奥行き推定を使用して、実世界物体の３次元位置および３次元速度を推定することによって、実世界物体を視覚的に追跡するための方法であって、
（１．１）時間ｔに計測されるカメラ画像を撮影するステップと、
（１．２）時間ｔの入力特徴を得るために一連のキューを使用して、時間ｔのカメラ画像のうちの追跡される物体が予想される部分領域を前処理するステップと、
（１．３）時間ｔの視覚的入力の領域であって、追加的アルゴリズムまたはユーザインタラクションのいずれかにより得られた領域を、外部手段を用いて指示することによって、時間ｔの入力特徴を使用してトラッカのテンプレートを初期化するステップと、
（１．４）時間ｔ＋ｄｔに次のカメラ画像を撮影するステップと、
（１．５）時間ｔ＋ｄｔのカメラ画像の部分領域をステップ１．２と同様に前処理するステップと、
（１．６）時間ｔおよびｔ＋ｄｔの入力特徴に２次元トラッカを使用して、カメラ画像の２次元座標における物体の見え方の２次元位置および２次元速度の推定値を得るステップと、
（１．７）奥行き変化を概算するために、２つの連続時間ステップｔおよびｔ＋ｄｔで測定された、追加的キューからの時間ｔの物体の奥行きの推定値を使用するステップと、
（１．８）時間ｔおよびｔ＋ｄｔのカメラ画像および／または選択された入力特徴に２次元変換推定を使用して、追跡される物体のスケール／サイズの相対的変化を抽出するステップと、
（１．９）ステップ１．７からの物体の奥行きおよび奥行き変化の概算推定を、ステップ１．８からのスケール／サイズの変化と結合して、物体の奥行き推定を改善するステップと、
（１．１０）カメラ座標におけるステップ１．６からの追跡される物体の２次元位置および２次元速度を、ステップ１．９からの奥行きおよび奥行き変化の推定と結合し、カメラ位置決め情報を使用することによってそれをグローバル３次元座標に変換して、追跡される物体のグローバル座標を得るステップと、
（１．１１）３次元位置を使用して、物体の大まかな物理的サイズを計算するステップと、
（１．１２）停止基準が満たされるまで、物体を追跡しながらステップ１．４～１．１１を繰り返すステップと、
を含む方法。
【請求項２】
（２．１）グローバル空間におけるカメラの位置および向きの変化を考慮に入れて、カメラおよび／または前記カメラが搭載されたプラットフォームの動きを補償するステップ、
をさらに含む、請求項１に記載の方法。
【請求項３】
（３．１）ステップ１．６および１．７による物体状態の推定、および／またはステップ１．９による結合に確率論的方法を使用することによって、不確実性を考慮に入れる、請求項１または２に記載の方法。
【請求項４】
（４．１）ステップ１．７からの奥行きおよび奥行き変化／奥行き速度の概算推定が、結果を時間で積分することによって増分的に行なわれる、請求項１ないし３のいずれか１項に記載の方法。
【請求項５】
（５．１）単一の概算奥行き推定の代わりに、異なるキューおよび／または測定技術に基づく一連の奥行き推定が使用され、次いでそれらが再び、ステップ１．９の場合と同様

にスケール／サイズの変化推定で２次元変化と結合される、請求項１ないし４のいずれか１項に記載の方法。
【請求項６】
（６．１）ステップ１．８からの２次元変換の推定が、ステップ１．７からの予想される奥行き変化／奥行き速度を考慮に入れることによって行なわれ、すなわち、奥行きの増加／減少によって生じる予想されるサイズの増減が、変換探索手順で考慮され、
（６．２）ステップ１．７からの物体の奥行き推定が、ステップ１．１１で計算された物理的サイズから導出される予想奥行き、およびステップ１．８からの追跡対象物体のスケール／サイズの予想変化に関する事前の情報を使用することによって行なわれる、
という意味で、２つの推定ステップ１．７および１．８が相互に影響を及ぼす、請求項１ないし５に記載の方法。
【請求項７】
（７．１）物体の状態パラメータのより高次の導関数に同じ原理が適用される、請求項１ないし６に記載の方法。
【請求項８】
カメラ手段の位置および向きを適応させるためのアクチュエータを制御することによって、カメラ手段の入力視野内の追跡される物体の位置および向きを制御するビジュアルサーボユニットに、ステップ１．１２の結果が転送される、請求項１ないし７のいずれか１項に記載の方法。
【請求項９】
請求項１ないし８のいずれか１項に記載の方法を実行するようにプログラムされたコンピューティング手段に信号を供給するカメラ手段を有する追跡装置。
【請求項１０】
請求項９に記載の追跡装置を具備したヒューマノイドロボット。
【請求項１１】
請求項９に記載の追跡装置を具備した自動車。
【請求項１２】
コンピューティング装置で実行したときに請求項１ないし８のいずれか１項に記載の方法を実現する、コンピュータソフトウェアプログラム製品。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、カメラシステム、２次元画像形成、およびカメラからの物体の距離の異なる測定値の組合せを用いて、経時的に実世界物体の動的状態を推定するための方法を記載する。本発明はまた、カメラ手段およびプログラムされたコンピューティング手段を有する追跡装置にも関する。
【背景技術】
【０００２】
技術的システムにとって、視覚的追跡は、動的環境で物体を解析するために必要な１つの重要な特徴であり、この数十年の集中的な研究の対象であり、例えば監視、衝突防止、および軌跡評価の分野における用途を導いてきた。
【０００３】
Toward Robot Learning of Tool Manipulation from Human Demonstration, Aaron Edsinger and Charles C. Kemp 1 Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology Cambridge, Massachusetts（特に図１および１０を参照されたい）に詳述されている通り、ロボット分野では、カメラ手段の位置および向きを適応させるためのアクチュエータを制御することによって、カメラ手段の入力視野内の追跡される物体の位置および向きを制御するビジュアルサーボユニットに、視覚的追跡の結果を転送することができる。
【０００４】

ビジュアルサーボは、１つ以上のカメラおよびコンピュータビジョンシステムを使用して、ワーク（追跡される物体）に対するロボットのエンドエフェクタ（マニピュレータ）の位置を制御することを含む。

【０００５】

技術的追跡システムの主な問題は、それらがしばしば、追跡されている物体の正確な内部モデルを必要とすることにある。一例として、交通場面では、これらは例えば自動車の３次元モデル（それらの正確な物理的大きさについての知識を含め、ボリュメトリックまたはサーフェスモデルのいずれか）とすることができ、それらは次いで、カメラシステムによって感受される刺激と一致するように適合される。代替的に、多くの技術的追跡システムは、物体の色のような物体の見え方に関する特定の知識によって、物体を見つける。しかし、一般的な場合、追跡すべき物体は事前には分からないので、正確な３次元モデルまたは他の特定的な知識は利用できない。この場合、追跡システムは、幾つかの異なるキューおよび測定値の組合せを使用する、物体の３次元位置および３次元速度の推定に頼らなければならない。

【０００６】

先行技術

視覚的追跡とは、カメラ手段によって供給される信号を使用して、経時的に実世界物体を、その動的パラメータ（位置、速度等）およびカメラにおけるその２次元見え方が変化するにも関わらず、視覚的に識別しかつ追従する能力である。カメラの見え方は、実世界３次元物体の視覚的特性の２次元スクリーンへの透視投影の性格を成すので、本質的に２次元である。それは、（外部光スペクトル、光源の位置の変更、反射率、および日陰効果のような）変わりやすい外部条件によって生じる様々な表面効果のため、物体の変形のような内部特性のため、または単に物体が回転して奥行きの位置が変わったために、かなり変化することがあるので、透視投影は、捕捉された画像における物体の異なる２次元見え方を導く。

【０００７】

さらに詳しくは、視覚的追跡は通常、経時的な物体の動的制約付き探索を定義する。これは、物体の動的状態（その位置、速度等）およびさらなる変換パラメータの推定を含み、両タイプの動的パラメータは通常、内部に格納された見え方モデルと、物体の現在の実際の見え方をもたらす刺激との間の一致を最大化しようと試みる、対応探索によって得られる。最大化は、物体の仮説的動的パラメータに応じて内部見え方モデルを変化させることによって、実現される。次いで最良の一致は、物体のさらなる追跡のために使用される真のパラメータの新しい推定値を得るために、処理される。追跡メカニズムの研究は、非特許文献１に見ることができる。

【０００８】

視覚的追跡のための先行技術として、以下が挙げられる。

【０００９】

１．幾つかの公知の追跡システムは、２次元「テンプレート」を使用して、物体パラメータとして例えばその２次元位置、速度、および加速度を推定して、純粋に２次元ベースの追跡を取り扱う。例として、テンプレートと入力との間のユークリッド差分費用関数を使用する視覚入力における物体の相関ベースの探索、ヒストグラムベースの費用関数（非特許文献２、３）を使用する平均場技術、およびユークリッド費用関数の線形化バージョンを構成する差分法（例えばＬｕｃａｓ－Ｋａｎａｄｅまたは「ＫＬＴ」アルゴリズム）のような特別なアルゴリズムが挙げられる。以下で、これらの技術のいずれかに従って働くモジュールを「２次元トラッカ」と呼ぶ。

【００１０】

２．テンプレートマッチング技術を使用して、回転、スケーリング、および剪断を含む幾何学的２次元変換のような、より複雑な追跡の物体パラメータを推定することができる。例として再び、アフィン変換の推定のために特殊な変形を施した、「ＫＬＴ」アルゴリズムが挙げられる（非特許文献４～６）。以下で、この技術を「２次元変換推定」と呼ぶ

。
【００１１】
３．奥行きは通常、別個のキューから得られる追加的な「測定値」として含まれる。現状技術では、視差の計算を可能にする両眼／立体視システムが使用され、その結果として、追跡される物体の奥行き推定が得られる（非特許文献７）。奥行きは、物体の状態に付加されるが、自律的に実行される２次元ベースの追跡に影響を及ぼさない。視差ベースの奥行き計算は、信頼性の点でかなりの限界を有し、ベースライン長（２つのカメラ間の水平距離）によって制限される狭い奥行き範囲でのみ有効であるので、例えば人間は、数メートルの範囲でしかそれを使用することができない。以下で、この技術を「視差ベースの奥行き測定」と呼ぶ。
【００１２】
４．２次元トラッカの代わりに、３次元トラッカは時々、物体の正確な内部３次元モデルを使用する（物体を他の手段で測定するか、または事前に承知していなければならない）（非特許文献８、９）。この場合、内部３次元モデルの変換投影バージョンとカメラ入力との正確な一致を見出すことができれば、網膜上のそのサイズから、その奥行きに関する結論を引き出すことが可能である。
【００１３】
５．別のタイプの両眼３次元トラッカは、内部状態として３次元座標を直接使用して始動し、両方のカメラで特定の物体の見え方を見つけようとする。この場合、２次元見え方一致計算および奥行き推定は、３次元座標、および左右のカメラの２次元座標へのそれらの投影を介して、自動的に結合される。これらのトラッカはしばしば、（特殊な事例として、カルマンフィルタを含む）動的ベイジアンネットワークを使用して実現されるので、推定が経時的に得られかつ改善される。
【００１４】
６．公知のマルチキュートラッカは、同一タイプのパラメータの推定のために複数のキューを統合し、それらの信頼性に従ってそれらを組み合わせるか、または最も信頼できるキューを選択して、これらのみに基づいて推定を行なう。
【００１５】
７．一般的に、動的ベイジアン推定器はまさしく、追跡中に生じる状態変数の時間的推定および統合のための研究分野である。我々は、その変形の粒子フィルタおよびカルマンフィルタを含むこの先行技術を公知であると考える（非特許文献１０～１２）。
【００１６】
８．視界のための頑健かつ高密度の３次元信号を提供する、例えば飛行時間信号を利用した奥行き検知カメラ技術（特許文献１を参照されたい）を使用する、多数の視覚的追跡システムが存在する。通常、これらのシステムは、３次元データに頼って物体を検出し追跡する。すなわち、物体は奥行きデータを用いて「切り取られる」。しかし、本発明で追求する方法は標準的な可視カメラおよび見え方ベースの追跡に依存するので、ａ）追跡される物体はその３次元データからセグメント化可能である必要が無く、かつｂ）特殊な検知ハードウェアが不要である。それにも関わらず、そのような検知技術からのデータを、まさしく核心的な意味で、追加的奥行き測定値として我々のシステムに組み込むことができる。
【先行技術文献】
【特許文献】
【００１７】
【特許文献１】ＷＯ２００４／１０７２６６Ａ１
【非特許文献】
【００１８】
【非特許文献１】Yilmaz, A., Javed, O., Shah, M. “Object tracking: A survey”. ACM Comput. Surv. 38(4) (2006) 13
【非特許文献２】Comaniciu, D., Ramesh, V., Meer, P. “Real-time tracking of non-

rigid objects using mean-shift”. Computer Vision and Pattern Recognition, 02:21 42, 2000
【非特許文献３】Comaniciu, V., Meer, P. “Kernel-based object tracking”, 2003
【非特許文献４】Lucas, B.D., Kanade, T. “An iterative image registration techni que with an application to stereo vision“. In International Joint Conference on Artificial Intelligence (IJCAI81), pages 674-679, 1981.
【非特許文献５】Tomasi, C., Kanade, T. “Detection and tracking of point feature s”. Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.
【非特許文献６】Shi, J., Tomasi, C. “Good features to track”. In IEEE Conferen ce on Computer Vision and Pattern Recognition (CVPR’94), pages 593-600, Seattle , June 1994.
【非特許文献７】Qian, N. “Binocular disparity and the perception of depth”. Ne uron 18(3): 359-368, March 1997.
【非特許文献８】Lou, J., Tan, T., Hu, W., Yang, H., Maybank, S.J. “3-D Model-Ba sed Vehicle Tracking”, IEEE Transactions on Image Processing, Volume 14, pp. 15 61-1569, Oct. 2005.
【非特許文献９】Krahnstoever, N., Sharma, R. “Appearance Mangement and Cue Fusi on for 3D Model-Based Tracking”, IEEE Computer Society Conference on Computer V ision and Pattern Recognition, Volume 2, pp. 249-254, 2003.
【非特許文献１０】Giebel, J., Gavrila, D., Schnorr, C. “A Bayesian for multicue 3D object tracking“. In Proceedings of European Conference on Computer Vision, 2004.
【非特許文献１１】Ristic, B. “Beyond the Kalman Filter: Particle Filters for Tr acking Applications”. Artech House Publishers, February 2004.
【非特許文献１２】Arulampalam, S., Maskell, S., Gordon, N.J., Clapp, T. “A Tuto rial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking” . In IEEE Transactions of Signal Processing, Vol. 50(2), pages 174-188, February 2002.

【発明の概要】
【発明が解決しようとする課題】
【００１９】
本発明は、事前に正確な物体モデルを必要としないが、それにも関わらず、物体の２次元見え方の解析および奥行き推定のための幾つかの視覚的キューの統合に依存して、物体の３次元位置および３次元速度を推定することのできる視覚的追跡システムを目標とする。
【００２０】
この能力は、独立請求項の特徴によって達成される。従属請求項は、本発明の中心的概念をさらに展開する。
【課題を解決するための手段】
【００２１】
本発明の第１態様は、２次元見え方ヒントおよびマルチキュー奥行き推定を使用して、任意の実世界物体の３次元位置および３次元速度を推定することによって、該実世界物体を追跡するための方法に関する。該方法は、以下のステップを含む。
【００２２】
（１．１）時間ｔに計測されるカメラ画像を撮影するステップ。
【００２３】
（１．２）時間ｔの入力特徴を得るために一連のキューを使用して、時間ｔのカメラ画像のうちの追跡される物体が予想される部分領域を前処理するステップ。
【００２４】
（１．３）時間ｔの視覚的入力の領域であって、（セグメント化アルゴリズムのような

）追加的アルゴリズム、または（カメラによって送達される画像の位置および領域を指示する）ユーザインタラクションのいずれかにより得られた領域を、外部手段を用いて指示することによって、時間ｔの入力特徴を使用してトラッカのテンプレートを初期化するステップ。この時間ステップで、選択された領域は重み付けマスクの形で格納され、物体の内部状態は、（例えば、そのセントロイド、その重心、または重み付けマスクを用いて積分されたその平均位置を使用して）該領域から導出される位置に初期化される。

【００２５】

（１．４）時間ｔ＋ｄｔに次のカメラ画像を撮影するステップ。

【００２６】

（１．５）時間ｔ＋ｄｔのカメラ画像の部分領域をステップ１．２と同様に前処理するステップ。

【００２７】

（１．６）時間ｔおよびｔ＋ｄｔの入力特徴に２次元トラッカを使用して、カメラ画像の２次元座標における物体の見え方の２次元位置および２次元速度の推定値を得るステップ。

【００２８】

（１．７）視覚をベースにすることができるが、必ずしもそうする必要の無い追加的キューから、時間ｔの物体の奥行き（カメラシステムからのその距離）の概算推定を使用するステップ。２つの連続的時間ステップｔおよびｔ＋ｄｔで計測され、これは奥行き変化の概算として役立つ。実際には、奥行き計測は、双眼カメラシステムの場合のように、例えば第２カメラからの入力が関係するかもしれない。

【００２９】

（１．８）時間ｔおよびｔ＋ｄｔのカメラ画像および／または選択された入力特徴に２次元変換推定を使用して、追跡される物体のスケール／サイズの相対的変化を抽出するステップ。

【００３０】

（１．９）剛性物体の場合、サイズの縮小または拡大がそれぞれ物体の奥行きの増加または減少のヒントになるという意味で、ステップ１．７からの物体の奥行きおよび奥行き変化の概算推定を、ステップ１．８からのスケール／サイズの変化と結合して、物体の奥行き推定を改善するステップ。

【００３１】

（１．１０）（カメラ座標における）ステップ１．６からの追跡される物体の２次元位置および２次元速度を、ステップ１．９からの奥行きおよび奥行き変化の推定と結合し、カメラ位置決め情報を使用することによってそれをグローバル３次元座標に変換して、追跡される物体のグローバル座標を得るステップ。

【００３２】

（１．１１）３次元位置を使用して、物体の大まかな物理的サイズを計算するステップ。

【００３３】

（１．１２）物体を見失うまで（それは例えば入力とテンプレートとの間のマッチングから何らかの信頼性基準によって検出することができる）、物体を追跡しながらステップ１．４～１．１１を繰り返すステップ。

【００３４】

（１．１３）再びステップ１．１から始め、同一物体または新しい物体を追跡する。

【００３５】

該方法はさらに、

（２．１）グローバル空間におけるカメラの位置および向きの変化を考慮に入れて、カメラおよび／またはそれが搭載されたプラットフォームの動きを補償するステップ、

を含む。

【００３６】

ステップ１．６および１．７による物体の推定、および／またはステップ１．９による結合に確率論的方法を使用することによって、不確実性を考慮に入れることができる。

【００３７】

ステップ１．７からの奥行きおよび奥行き変化／奥行き速度の概算推定は、例えば動的ベイジアンフィルタ／カルマンフィルタ等を使用するときのように、結果を時間で積分することによって増分的に行なうことができる。

【００３８】

単一の概算奥行き推定の代わりに、異なるキューおよび／または測定技術に基づく一連の奥行き推定を使用することができ、次いでそれらは再び、ステップ１．９の場合と同様に、スケール／サイズの変化推定で２次元変化と結合される。

【００３９】

２つの推定ステップ１．７および１．８は、次の意味で、相互に影響を及ぼすことがある。

ａ）ステップ１．８からの２次元変換の推定は、ステップ１．７からの予想される奥行き変化／奥行き速度を考慮に入れることによって行なわれる。すなわち、奥行きの増加／減少によって生じる予想されるサイズの増減が、変換探索手順で考慮される。

ｂ）ステップ１．７からの物体の奥行き推定は、ステップ１．１１で計算された物理的サイズから導出される予想奥行き、およびステップ１．８からの追跡対象物体のスケール／サイズの予想変化に関する事前の情報を使用することによって行なわれる。

【００４０】

物体の状態パラメータのより高次の導関数に同じ原理を適用することができるので、例えば物体の位置またはサイズまたは向きの加速度が推定され、追跡手順に使用される。

【００４１】

本発明のさらなる特徴、特性、および利点は、本発明の好適な実施形態の以下の詳細な説明を、添付する図面の図と併せて読んだときに、当業者には明らかになるであろう。

【図面の簡単な説明】

【００４２】

【図１】本発明を実現するためのシステムの概要を示す。

【図２】物体の奥行き、物体の物理的サイズ、およびその２次元見え方サイズの間の数学的関係の図解を示す。

【発明を実施するための形態】

【００４３】

図１は、本発明を実現するためのシステムの概要を示す。

【００４４】

該システムは、幾つかのモジュールを備える。

【００４５】

１．物体の２次元見え方に基づいて働き、その位置および速度を推定する２次元トラッカモジュール。これは、追跡される物体の正確な事前のモデルを必要とすることなく行なわれる。代わりに、追跡テンプレート（すなわち、今から追跡すべき場面の原型的部分）を、第１ステップとして入力画像から直接抽出することができる。これは、物体の見え方および物体クラス／タイプに対する事前の制限無く、一般的な任意の物体を追跡することを可能にする。

【００４６】

２．そのスケールおよびスケール変化率を含め、テンプレートと入力との間の最良の一致を見出すことを可能にする、幾つかの変換パラメータを推定する、幾何学的２次元変換推定モジュール。例として、２つの追跡時間ステップ間のスケール変化が０．１である場合、それは、追跡される物体の２次元見え方（すなわち、カメラシステムによって送達されるときのその見え方）は、サイズが１０％増大することを意味する。

【００４７】

３．物体の奥行き（すなわち、カメラシステムから物体までの距離）の独立測定値を提

供する１つ以上のモジュール。しかし、これらの測定値はあまり信頼できないかもしれない。我々の例示的な場合では、ａ）左右のカメラ間のローカルパッチの比較に基づく高密度双眼視差ベースの奥行き測定システム、およびｂ）左右のカメラ画像内で物体を見つけ出そうとし、２つの一致間のずれから追跡される物体の単一の奥行き値を計算する、よりグローバルに作用する双眼システムを使用する。

【００４８】

４．座標をカメラ座標系からグローバル座標系に変換しかつ元に戻すための手段。この場合、各ポイントの「完全な」３次元カメラ座標は、２次元カメラ座標プラス奥行き値から構成される。カメラ画像は、一種の「ピンホールカメラ」として取り扱うことができるように、レンズ歪みを考慮するように修正することができると想定し、グローバル空間におけるカメラの位置および向きが分かると、幾何学的考察を介して座標変換を容易に求めることができる。その場合、１つの座標系から他の座標系への完全な変換は、平行移動、直交投影、および透視投影の観点から説明することができる。

【００４９】

５．システムと共にカメラシステムの動きを定量化することができるように、グローバル空間におけるカメラの位置および向きの変化を測定または推定する手段。これは、例えばビジョンシステムを搭載した動く自動車またはロボットのような、カメラを搭載したまま自律的に移動するプラットフォームの場合に特に有利である。

【００５０】

システムは、場合により可動であるプラットフォームに取り付けられたカメラから、ｗ＊ｈ画素のサイズの２次元入力画像を受信する。グローバルに固定された座標系に対するプラットフォームの位置および向き、ならびにプラットフォームに対するカメラ（または、物体奥行き測定が双眼視差によって供給される場合、２つのカメラ）の位置および向きは、いつでも大まかに分かっているものと想定される。これにより、グローバル３次元座標系とカメラに固定されそれとアライメントさせた３次元座標系との間で座標の変換が可能である。

【００５１】

さらに、カメラの３次元座標からカメラの２次元入力座標への投影およびその逆が分かっているものと想定される。その結果、グローバル３次元座標系と２次元カメラ入力座標との間で、座標をいつでも変換することができる。

【００５２】

次に、本発明に係る処理について説明する。

【００５３】

第１ステップとして、２次元カメラ入力は、一連のｎ個のキューを使用して前処理され、ｎ個１組の画像を導く。キューは、向き、コントラスト、または色抽出フィルタを入力画像に適用するような、単純な前処理方法によって得ることができるが、その期待される色のような、特定の物体に特異的な、より洗練された指標をも含むことができる。キューの厳密な選択は、それらが前処理されたカメラ画像を用いて物体を追跡することを可能にする充分な情報を含む限り、本発明の関連事項ではない。理想的な場合では、１つのキューからの追跡障害が他のキューによって補償され、プロセス全体が入力の変動に対して頑健となるように、様々なキューが、物体に関して非相関的な情報を伝達する。以下のステップのほとんどは、前処理された入力画像から抽出されたｎ個のキューを使用し、以下ではそれらを「ｎ個の入力特徴」と呼ぶ。

【００５４】

処理資源を節約するために、前処理を２次元カメラ視野の制限された部分領域に制限することが可能である。部分領域は、例えば、追跡すべき物体の推定される位置、サイズ、および速度から決定することができる。ひとたび追跡が開始されると、これらのパラメータは全て、追跡システムによって連続的に推定されるので、それらは容易に利用可能であり、各時間ステップで部分領域をそれに応じて調整することができる。

【００５５】

まず、手始めに、システムは、追跡すべき物体がどのように見えるか、おおよその概念を必要とする。この目的のために、システムに、入力特徴と同一空間で作用する２次元見え方テンプレートを供給しなければならない。実際には、これは、システムがそのようなテンプレートをユーザインタラクションもしくはここに記載しない他の（以前に記録された見え方テンプレートを供給することのできるメモリモジュールのような）モジュールから得ること、または再びユーザインタラクションからまたは他のモジュールから位置およびエリア情報を得て、現在の入力特徴を使用して、独力で２次元見え方テンプレートを抽出することを意味する。詳しくは、テンプレートを抽出するには、直接供給されるかまたは２次元カメラ入力座標に変換することのできる、位置およびエリアの指示を必要とする。システム内で物体の２次元見え方を説明するために使用される物体の内部状態は、テンプレート、２次元カメラ座標における位置、および指示されたエリアから得られる重み付けマスクから構成される。最終的に、例えば位置が３次元座標で供給された場合、想定される速度、サイズ等のような補助情報のみならず、３次元位置も物体の状態の一部となる。

【００５６】

ひとたびトラッカが初期化されると、２次元見え方テンプレートと入力特徴との間の一致が大きい視覚的場面の部分を見つけることによって、その後の画像で物体を探索することができる。これは２次元トラッカによって達成され、それは、我々の特別な場合では、増分的にかつ統計的に動作する最先端のマルチキュー動的ベイジアントラッカとして実現される。このモジュールから得られる結果は、追跡される物体が現在何か特定の２次元カメラ入力位置および速度を有している確率を示す、確率マップである。

【００５７】

物体の２次元見え方が経時的にかなり変化する場合、２次元見え方テンプレートおよび／または物体特定的なキューのパラメータを再調整する、キューおよびテンプレート適応ステップを、追跡手順に組み込むことが有用である。

【００５８】

さらなるステップとして、物体とカメラとの間の３次元距離を意味する、奥行きがここで推定される。本発明の中心的概念は、全く異なるキューに作用しかつ各々の単一推定の弱点を相互に補い合うように相補的に機能する、２つの特定の（以下で説明する）奥行き推定を結合するというものである。１）それらの１つは、双眼視を利用するような従来の技術を使用した、直接奥行き推定である。本質的に、この直接奥行き推定は、相互にすでに統合された幾つかの異なる方法の組合せとすることができる。２）第２の技術は、単一のカメラの物体の２次元見え方のサイズの変化を観察することから抽出できる奥行き変化推定である。基本原理は、カメラまでの距離が増加または減少する物体が、２次元見え方のサイズの減少または増加をそれぞれ生じるということである。物体奥行きおよび奥行き速度が分かると、物体の２次元見え方の予想サイズ変化を推定することができる。逆に、物体奥行きおよびその２次元見え方サイズの変化が分かると、奥行き速度を推定することができる。したがって２つの奥行き推定（すなわち、直接推定および２次元サイズ変化の推定に基づくもの）は、密接に結び付けられ、これら２つの特性の組合せを利用する方法で３次元追跡システムを設計することは合理的である。そのようなシステムは、物体自体の正確な３次元モデル無しで、主に物体の２次元見え方および２次元見え方変化に頼って、任意の物体を安定的に容易に３次元追跡することが可能である。

【００５９】

特に、双眼技術と組み合わせて、２つの異なる奥行き推定法を結合することにより、３次元物体追跡システムの利点がもたらされる。双眼システムは、カメラのベースライン長に応じて、カメラからの近接距離でよく機能する。しかし、追跡される物体のサイズ変化の観察による奥行き推定は、より大きい範囲の距離に対してよく機能することができる。したがって両方の方法は相互にブートストラップすることができる。例として、その奥行きが近接距離範囲で正確に測定される物体は、それが双眼システムの正確な３次元推定の範囲外に移動したときには、そのサイズ変化を奥行き測定して追跡することができる。

【００６０】
物体の２次元見え方のサイズ変化は、このシステムで、先行技術の要点２に記載した、２次元変換推定と呼ばれるテンプレートマッチング手順を使用して、抽出される。我々の場合では、システムは、２次元見え方テンプレートと入力画像との間の最良の一致を提供するアフィン変換Ａを探索する。各時間ステップt_kで、現在の変換状態Ａ（t_k）は、「現在追跡される物体が、特定の量だけ変換された、例えば５度回転されかつ１０％スケーリングされた、２次元見え方テンプレートのように見える」という意味で、テンプレートに対する追跡される物体の現在の２次元見え方の最良の表現を表わす。次いで次の時間ステップt_{k+1}の変換状態Ａ（t_{k+1}）は、変換状態Ａ（t_k）を未来に波及させ、かつそこから新しいテンプレートマッチングをベースとする変換推定探索を開始することによって、推定される。各アフィン変換状態から、サイズλは、次のようにそれをスケーリングおよび回転の合成として近似することによって抽出される。
【数１】

$$A(t_k)=\begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \approx \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \bullet \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \qquad (1)$$

【００６１】
これは、スケールλが、変換状態行列の決定から直接計算することができることを意味する。次いで、２つの連続時間ステップからの結果を減算して、２次元見え方のサイズ変化を定量化する変換変化、例えば２つの時間ステップ間のスケール変換変化Δλを計算することができる。
【００６２】
双眼入力に基づく直接奥行き推定は、我々の場合では、標準視差ベースの高密度（すなわち画素単位）奥行き測定（先行技術の要点３を参照されたい）の後に、物体重み付けマスクを用いて、（例えば、高密度奥行き測定とマスクの空間積分によって）追跡される物体全体の単一のおおよその奥行きを抽出するものであった。再びこれらの測定値を、動的ベイジアン推定器を用いて時間で積分した。我々がシステムに組み込んだ追加の第２の双眼奥行き測定は、第１カメラから抽出された２次元見え方テンプレートを、例えば相互相関または先行技術の要点２に記載されたテンプレートマッチング技術を用いて、第２カメラで直接探索するものである。左側のカメラでテンプレートが見られる位置と右側のカメラで見られる位置との間の相対ずれから、奥行き測定のベースとしても使用することのできる視差が抽出される。
【００６３】
直接奥行き測定は、しばしば信頼できない物体の奥行きに関して、幾つかのヒントを提供する。連続時間ステップからの奥行きを使用して、観察者／カメラプラットフォームに対して移動する物体の奥行きの速度を抽出することもできる。カメラプラットフォーム自体が移動しており、かつこの動きのパラメータが既知である場合には、例えばカメラ位置および動きを物体パラメータから減算して、絶対座標の物体パラメータを得ることによって、この作用を補償することができる。
【００６４】
直接奥行き測定に加えて、システムは説明した通り、２次元見え方サイズ変化測定値を送達する。略剛性の物体の場合、較正済みのカメラシステムでは、その物理的サイズλ_{ph}、その２次元見え方サイズλ、および奥行きｚの間の関係は、次の通りである（ｃは、カメラの焦点距離、２次元スクリーンサイズ、および追跡テンプレートサイズのような、幾つかのシステムパラメータを圧縮する定数である）。

【数２】

$$\frac{z}{c}=\frac{\lambda_{ph}}{\lambda} \qquad (2)$$

【００６５】

これは、２次元見え方サイズλおよび奥行きｚが相互に反比例することが予想されることを表わす。つまり、同一の物理的サイズの場合、カメラからより遠い／により近い物体はカメラシステムではより小さく／より大きく見える。

【００６６】

図２は、物体奥行き、物体の物理的サイズ、およびその２次元見え方サイズの間の数学的関係の図解を示す。

【００６７】

本発明に係るシステムでは、数式２は直接奥行き測定を、２次元見え方サイズ変化による奥行き推論と結合する。内部では、我々はとりわけ速度のような、パラメータλ$_{ph}$、λ、および｛ｘ，ｙ，ｚ｝を含む状態によって追跡される物体を表わす。直接奥行き測定は、各時間ステップでｚの新しい推定値をもたらす。２次元見え方変化測定は、各時間ステップで、λの新しい推定値をもたらす。２次元位置追跡は奥行きと共に、各時間ステップで物体のグローバル位置｛ｘ，ｙ，ｚ｝の新しい推定値をもたらす。物理的サイズλ$_{ph}$は、他の感覚測定、または代替的に特定の物体の物理的サイズに関する事前の知識、または対話する人間からの管理入力等のような、その状態に関する追加ヒントを他のソースから受け取ることのできるシステムの内部パラメータである。

【００６８】

ここで追跡システムのタスクは、現在の状態パラメータλ$_{ph}$（ｔ$_{k}$）、λ（ｔ$_{k}$）、および｛ｘ，ｙ，ｚ｝（ｔ$_{k}$）を取り、それらを使用して、状態パラメータの何らかの動的モデル（例えば物体の物理的サイズが一定であり、物体が一定の奥行き速度で移動しており、かつ２次元見え方サイズが数式２に従って変化するようなモデル）に基づいて、次の時間ステップの予想状態パラメータ

【数３】

$$\hat{\lambda}_{ph}(t_{k+1}),\ \hat{\lambda}(t_{k+1})$$

および

【数４】

$$\{\hat{x},\hat{y},\hat{z}\}(t_{k+1})$$

を推定し、これを（２次元サイズ推定からの）λ、（直接奥行き推定からの）ｚ、および（２次元位置推定からの）｛ｘ，ｙ｝の新しい測定推定値と結合して、全て数式２の制約の下で新しい状態パラメータλ$_{ph}$（ｔ$_{k+1}$）、λ（ｔ$_{k+1}$）、および｛ｘ，ｙ，ｚ｝（ｔ$_{k+1}$）の更新推定値を得ることである。（図１で、新しい測定推定値は、それらを「真」の測定値と区別するために、表記法

【数５】

$$\bar{\lambda},\ \bar{z}$$

および

【数６】

$$\{\bar{x},\bar{y}\}$$

を受け取る。）

【００６９】

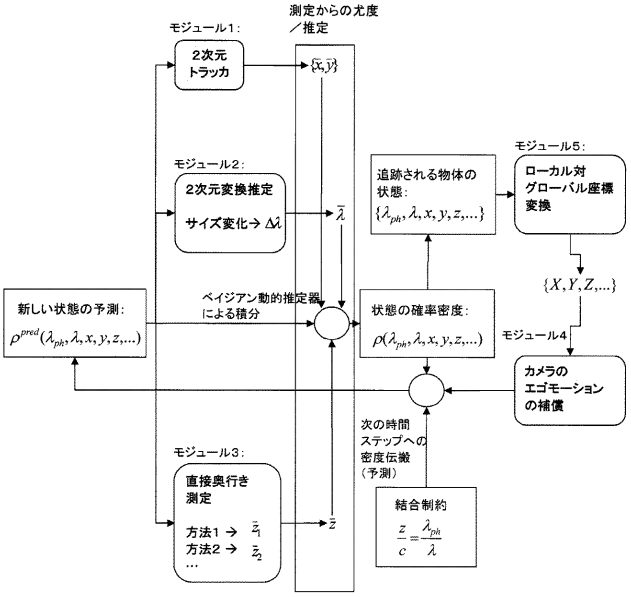
これを行なうための直接的方法は、少なくとも考慮する状態パラメータ

【数７】

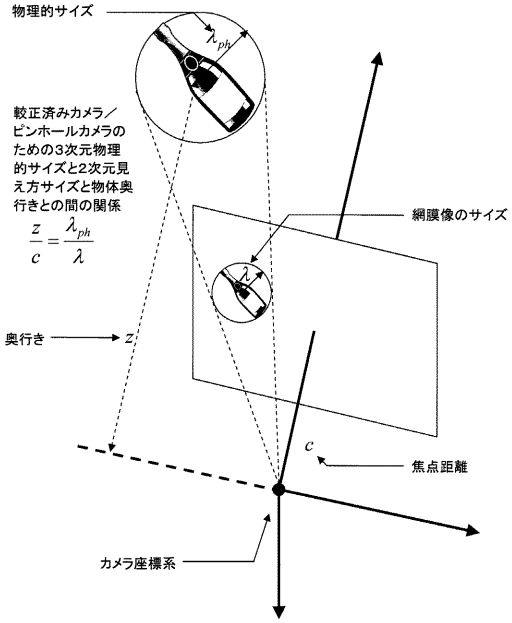
$$\rho(\lambda_{ph},\lambda,x,y,z,...)$$

を含む確率密度に対し、動的ベイジアン推定器／再帰的ベイジアンフィルタ／確率動力学的推定器の予測－確認枠組のような確率論的方法を使用し、過去の全ての直接奥行き測定および２次元見え方サイズ変化測定を前提として、経時的にそれを改善して、状態パラメータの最良の推定を得ることである。この枠組では、新しい測定推定値は、確率論的尤度に対応する。現在の確率密度から、追跡される物体の状況を最もよく記述する最確パラメータλ$_{ph}$、λ、およびｚを抽出することができる（例えば最大確率点を選択することによって；しかしこれを達成する異なる方法が存在する）。言うまでもなく、物体の（カメラ座標における）２次元位置および速度のようなさらなる物体パラメータが、見え方ベースの２次元追跡システムによって送達される（モジュール１）。次いでサイズおよび奥行きパラメータと共に、追跡される物体の３次元位置および速度を決定することができる（モジュール４）。さらに、３次元追跡とカメラの位置および向きの追跡の維持とを組み合わせることにより（モジュール５）、エゴモーション（ｅｇｏｍｏｔｉｏｎ）作用を補償することができるので、たとえカメラシステムがその位置を変えても、物体を確実に追跡することができる。特に、これは、ロボット／自動車自体が移動しながら、物体の信頼できる３次元追跡を行なうことが、視覚的場面の一貫した表現を構築するために必要な能力である、他の交通関与者の監視用のカメラを搭載した視覚的に案内されるロボットまたは自動車に関係する用途に、有意義である。

【図１】


測定からの尤度
／推定
モジュール1：
2次元
トラッカ
$\{\bar{x},\bar{y}\}$
モジュール2：
2次元変換推定
サイズ変化→ $\Delta\lambda$
$\bar{\lambda}$
追跡される物体の
状態：
$\{\lambda_{ph},\lambda,x,y,z,...\}$
モジュール5：
ローカル対
グローバル座標
変換
$\{X,Y,Z,...\}$
新しい状態の予測：
$\rho^{pred}(\lambda_{ph},\lambda,x,y,z,...)$
ベイジアン動的推定器
による積分
状態の確率密度：
$\rho(\lambda_{ph},\lambda,x,y,z,...)$
モジュール4
カメラの
エゴモーション
の補償
次の時間
ステップへの
密度伝搬
（予測）
モジュール3：
直接奥行き
測定
方法1 → $\bar{z}_1$
方法2 → $\bar{z}_2$
…
$\bar{z}$
結合制約
$\frac{z}{c}=\frac{\lambda_{ph}}{\lambda}$

【図２】


物理的サイズ
λ_{ph}
較正済みカメラ／
ピンホールカメラの
ための3次元物理
的サイズと2次元見
え方サイズと物体奥
行きとの間の関係
$\frac{z}{c}=\frac{\lambda_{ph}}{\lambda}$
網膜像のサイズ
λ
奥行き
z
c
焦点距離
カメラ座標系

フロントページの続き

(72)発明者　スヴェン・レバン
ドイツ国６３１７９　オーベルシャウゼン、ミュッルフェルドシュトラーセ　３
(72)発明者　ヴォルカー・ウィラート
ドイツ国６３５００　ゼリゲンシュタット、フランクフルター・シュトラーセ　５６
(72)発明者　チェン・チャン
ドイツ国６４２９５　ダルムシュタット、ルーデスハイマー・シュトラーセ　７５

Ｆターム(参考) 2F065 AA04 AA06 AA09 AA20 AA37 BB05 CC11 DD06 FF01 FF05
FF09 JJ03 JJ26 PP25 QQ17 QQ21 QQ25 QQ38
5L096 AA02 AA09 BA05 CA02 CA04 CA24 EA13 EA39 FA25 FA26
FA69 GA10 GA38 GA55 HA01 HA02 HA03 HA04 HA05 JA09

【外国語明細書】
1. TITLE OF INVENTION

Visually tracking an object in real world using 2D appearance and multicue depth estimations

2. DETAILED DESCRIPTION OF INVENTION

Technical Field

The present invention describes a method for the estimation of the dynamic state of a real-world object over time using a camera system, 2D image information and a combination of different measurements of the object's distance from the camera. The invention also relates to a tracking apparatus having camera means and programmed computing means.

Background of the Invention

For technical systems, visual tracking is one key feature necessary for analyzing objects in dynamic environments, and has been subject of intensive research during the last decades, leading to applications in the field of e.g. surveillance, collision avoidance and trajectory evaluation.

As it is explained in detail in Toward Robot Learning of Tool Manipulation from Human Demonstration, Aaron Edsinger and Charles C. Kemp 1 *Computer Science and Artificial Intelligence Laboratory Massachusetts Institute of Technology Cambridge, Massachusetts* (see particularly figures 1 and 10, in a robot application the result of the visual tracking can be forwarded to a visual servoing unit controlling the position and orientation of the tracked object in the input field of the camera means by

controlling actuators for adapting the position and orientation of the camera means.

Visual servoing may involve the use of one or more cameras and a computer vision system to control the position of a robot's end-effector (manipulator) relative to the workpiece (being the tracked object).

The main problem of technical tracking systems resides in that they often require accurate internal models of the objects that are being tracked. As an example, in a traffic scene, these can e.g. be 3D models (either volumetric or surface models, including knowledge about their exact physical dimensions) of cars, which are then fitted to match the stimulus as it is received by a camera system. Alternatively, many technical tracking systems find the objects by specific knowledge about an object's appearance, like its color. But in a general case, the objects that should be tracked are not known in advance, so that no accurate 3D model or other specific knowledge is available. In this case, tracking systems have to rely on estimations of an objects 3D position and 3D velocity using a combination of several different cues and measurements.

Prior art

Visual tracking is the capability to visually identify and follow, using a signal supplied by camera means, a real-world object over time despite that it is changing its dynamical parameters (position, velocity, etc.) and its 2D appearance on the camera. The camera appearance is

intrinsically 2D because it constitutes a perspective projection of the visual properties of the real-world 3D objects onto a 2D screen. It may change considerably due to different surface effects caused by variable external conditions (like external light spectrum, repositioning of light sources, reflectance and shading effects, etc.), internal properties like object deformations, or simply because the object rotates and repositions in depth, so that the perspective projection leads to a different 2D appearance of the object in the captured images.

More specifically, visual tracking usually defines a dynamically constrained search of the object over time. This includes the estimation of dynamic states of the object (its position, velocity, etc.) and further transformation parameters, with both types of dynamic parameters usually being gained by a correspondence search which tries to maximize the match between internally stored appearance models and the stimulus that provides the current, real appearance of an object. The maximization is realized by varying the internal appearance models depending on the objects hypothetic dynamic parameters. The best matches are then processed further to get a new estimate of the true parameters used for further tracking the object. A survey of tracking mechanisms can be found in [1].

Prior art techniques for visual tracking include the following:

1. Some known tracking systems deal with purely 2D-based tracking, estimating as object parameters e.g. its 2D position, velocity and acceleration using a 2D "template". Examples constitute special algorithms like correlation-based search of the objects in the visual input using an Euclidean difference cost function between the template and the input, mean-field techniques which use a histogram-based cost function [2,3], and differential methods which constitute a linearized version of the Euclidean cost function (e.g. Lucas-Kanade, or "KLT" algorithms). In the following, a module working according to any of these techniques will be referred to as "2D-tracker".
2. More complex object parameters for tracking, like geometrical 2D transformations including rotation, scaling and shearing can be estimated using template matching techniques. An example constitutes again the "KLT" algorithm, with special variants for the estimation of affine transformations [4-6]. In the following, this technique will be referred as "2D transformation estimation".

3. Depth is usually included as an additional "measurement", gained from a separate cue. State-of-the-art is to use a binocular/stereo vision system that allows for disparity computation, resulting in a depth estimate for the tracked object [7]. The depth is attached to the objects state but does not influence the 2D-based tracking, which runs autonomously. Disparity-based depth computation has considerable limitations in terms of reliability and only works well in a narrow depth range

limited by the baseline length (the horizontal displacement between the two cameras), so that e.g. humans are only able to use it in a range of a few meters. In the following this technique will be referred to as "disparity-based depth measurement".

4. Alternatively to 2D-trackers, 3D trackers sometimes use accurate internal 3D models of the object (which have to be measured by other means or known in advance) [8, 9]. In this case, if an accurate match of a transformed and projected version of the internal 3D model with the camera input can be found, its size on the retina allows drawing conclusions about its depth.

5. Another type of binocular 3D trackers starts directly using 3D coordinates as internal state, and tries to find a particular object appearance in both cameras. In this case, the 2D appearance match calculations and the depth estimation are coupled automatically via the 3D coordinates and their projection onto the 2D coordinates of the left and right cameras. These trackers are often implemented using Dynamic Bayesian Networks (including Kalman filters as a special case) so that the estimations are gained and improved over time.

6. Known multicue trackers integrate multiple cues for the estimation of the same type of parameters, either combining them according to their reliability or selecting the most reliable cue(s) and basing the estimation on these only.

7. Generally, Dynamic Bayesian Estimators are a very researched field for the temporal estimation and integration of state variables as it occurs during tracking. We assume this prior-art technique including its variants Particle Filter and Kalman Filter to be known [10-13].

8. There exist a number of visual tracking systems using depth-sensing camera technology, e.g. taking advantage of time-of-flight signals (see WO 2004/107266A1), which provide a robust and dense 3D signal for the view field. Usually, these systems rely on the 3D data to detect and track an object, i.e., the object is "cut-out" using the depth data. The approach pursued in this invention, however, relies on standard visual cameras and appearance-based tracking so that a) the tracked object does not need not be segmentable from its 3D data and b) no special sensing hardware is required. Nevertheless, data from such sensing technology could be integrated into our system as an additional depth measurement in the very sense of point.

Disclosure of the Invention

The present invention targets at a visual tracking system which does not need accurate object models in advance, but is nevertheless able to estimate an object's 3D position and 3D velocity relying on the analysis of an object's 2D appearance and the integration of several visual cues for depth estimation.

This capability is achieved by means of the features of the independent claims. The dependent claims develop further the central idea of the present invention.

A first aspect of the invention relates to a method for tracking arbitrary real-world objects by estimating their 3D position and 3D velocity using 2D appearance hints and multicue depth estimations. The method comprises the steps of:

(1.1.) Taking a camera image measured at time t,

(1.2.) Preprocessing a subregion of the camera image at time t where the object to be tracked is expected using a series of cues to get the input features at time t,

(1.3.) Using the input features at time t to initialize the tracker template by indication of a region of the visual input at time t using external means, with the region gained either from additional algorithms (like a segmentation algorithm) or by user interaction (indicating the position and region in the image delivered by the camera). At this time step, the selected region is also stored in form of a weighting mask and the internal state of the object gets initialized to a position derived from the region (e.g., using its centroid, its center of gravity or its averaged position integrated using the weighting mask),

(1.4.) Taking the next camera image at time t+dt,

(1.5.) Preprocessing a subregion of the camera image at time t+dt in the same way as in step 1.2,

(1.6.) Using a 2D-tracker on the input features at times t and t+dt to get estimates of the 2D position and 2D velocity of the objects appearance in camera image 2D coordinates,

(1.7.) Using an approximate estimation of the objects depth (its distance from the camera system) at time t from an additional cue, which can be but need not be visually based. Measured at two consecutive timesteps t and t+dt, this serves as an approximation for the depth change. In practice, the depth measurement may involve e.g. input from a second camera as in the case of a binocular camera system,

(1.8.) Using a 2D transformation estimation on the camera images and / or selected input features at times t and t+dt to extract the relative change of scale / size of the object that is being tracked.

(1.9.) Coupling the approximate estimation of the depth and the depth change of the object from step 1.7 with the change in scale / size from step 1.8 to improve the depth estimation for the object, in the sense that for rigid objects a reduction or an increase in size is a hint for an increase or decrease in depth, respectively, of an object,

(1.10.) Combining the 2D position and 2D velocity of the tracked object from step 1.6 (in camera coordinates) with the depth and depth change estimation from step 1.9 and converting it into global 3D coordinates by using camera positioning information, to get the global coordinates of the object that is being tracked,

(1.11.) Using the 3D position to calculate an approximate physical size of the object,

(1.12.) Iterating the steps 1.4 -1.11 while the object is being tracked, until the object is lost (which can be detected e.g. by some confidence criterion from the match between input and template), and

(1.13.) Starting again at step 1.1 to track the same or a new object.

The method may furthermore comprise the step of:

(2.1.) taking changes of the position and orientation of the camera in global space into account to compensate for motion of the camera and / or the platform where it is mounted on.

Uncertainties can be taken into account by using probabilistic methods for the estimation of the object states according to steps 1.6 and 1.7 and / or the coupling according to step 1.9.

The approximate depth and depth change / depth velocity estimation from step 1.7 may occur incrementally by integrating the results over time, like e.g. when using Dynamic Bayesian Filters / Kalman filters, etc.

Instead of a single approximate depth estimation a series of depth estimations based on different cues and / or measurement techniques can be used which are then again coupled with the 2D change in scale / size change estimation as in step 1.9.

The two estimation steps 1.7 and 1.8 may influence each other, in the sense that

a) the estimation of the 2D transformation from step 1.8 occurs by taking the expected depth change / depth velocity from step 1.7 into account, meaning that an expected, reduced or increased size caused by an increase / decrease in depth is considered in the transformation search procedure,

b) the estimation of the object's depth from step 1.7 occurs by using prior information about the expected depth derived from the physical size calculated in step 1.11 and the expected change of scale /size of the object being tracked from step 1.8.

The same principles may be applied to higher-order derivatives of object's state parameters, so that e.g. accelerations of an object's position or size or orientation are estimated and used for the tracking procedure.

Further features, properties and advantages of the present invention will become evident for the skilled person when reading the following detailed description of preferred embodiments of the invention, when taken in conjunction with the figures of the enclosed drawings.

Best Mode for Carrying out the Invention

Figure 1 shows an overview of a system for implementing the invention.

The system comprises several modules:

1. A 2D-tracker module which works based on the 2D appearance of an object to estimate its position and velocity. This occurs without the necessity of accurate prior models of the object that is being tracked; instead, the tracking templates (i.e., the prototypical parts of the scene that should be tracked from now on) can be directly extracted from the input image as a first step. This allows to track general, arbitrary objects without prior limitations on the object's appearance and object class/type.

2. A geometrical 2D transformation estimation module which estimates several transformation parameters that allow to find the best match between the template and the input, including its scale and the rate of scale change. As an example, a scale change of 0.1 between two tracking time steps would then mean that the 2D appearance of the object (i.e., its appearance as it is delivered by the camera system) that is being tracked has grown 10% in size.

3. One or several modules that provide independent measurements of the object's depth (i.e., the distance of the object from the camera system). These measurements may, however, be very unreliable. In our exemplary case, we use a) a dense, binocular disparity-based depth measurement system based on the comparison of local patches between the left and right cameras and b) a more globally working binocular system that tries to find the object in the left and the right camera

images and calculates a single depth value for a tracked object from the displacement between the two matches.

4. Means to transform coordinates from the camera coordinate system into a global coordinate system and back. In this case, the "complete" 3D camera coordinates for each point consist of the 2D camera coordinates plus a depth value. Assuming that the camera images can be rectified to account for lens distortions so that it can be treated as a sort of "pinhole-camera", and knowing the position and orientation of the camera in global space, one can easily find the coordinate transformations via geometrical considerations. The entire transformation from one coordinate system to the other can then be described in terms of a translation, an orthonormal projection and a perspective projection.

5. Means to measure or estimate changes of the position and orientation of the camera in global space, so that a motion of the system with the camera system can be quantified. This is of advantage especially for the case of autonomously moving platforms with mounted cameras, like e.g. moving cars or robots with vision systems.

The system receives its 2D input image with a size of w*h pixels from a camera that is attached to a platform which optionally may be movable. It is assumed that the position and orientation of the platform with respect to a globally anchored coordinate system, and the position and orientation of the camera (or the two cameras, in case that the objects depth measurement is supplied by

binocular disparity) with respect to the platform, are approximately known at any moment in time. This allows the conversion of coordinates back and forth between the global 3D coordinate systems and the 3D coordinate systems anchored at and aligned with the cameras.

Furthermore, it is assumed that the projections from the camera 3D coordinates to the camera 2D camera input coordinates and back are known. The consequence is that one can always convert coordinates back and forth between the global 3D coordinate system and the 2D camera input coordinates.

Now the processing according to the invention will be explained:

As a first step, the 2D camera input is preprocessed using a series of n cues, leading to a set of n images. The cues can be gained by simple preprocessing methods, like applying orientation, contrast or color extraction filters on the input image, but can also imply more sophisticated measures that are specific to a particular object, like its expected color. The exact choice of the cues is not of relevance for this invention as long as they contain sufficient information to be able to track an object using the preprocessed camera image. In the ideal case, the different cues convey uncorrelated information about the object, so that tracking deficits from one cue can be compensated by other cues, making the entire process robust against input variations. Most of the following steps use the n cues extracted from the preprocessed input

image, which will be called the "n input features" from here on.

To save processing resources it is possible to limit the preprocessing to a limited subregion of the 2D camera field. The subregion can e.g. be determined from an estimated position, size and velocity of the object that should be tracked. Once the tracking has started, all these parameters will be continuously estimated by the tracking system, so that they are readily available and at each timestep the subregion can be adjusted accordingly.

To start with, the system needs an approximate idea of how the object that should be tracked looks like. For this purpose, the system has to be supplied with a 2D appearance template, which works in the same space as the input features. In practice, this means that the system gets such a template from user interaction or other modules not described here (like memory modules that can supply a previously recorded appearance template), or that it gets a position and area information, again from user interaction or from other modules, and extracts the 2D appearance template by itself using the current input features. In detail, extracting the template requires indication of a position and an area that is supplied directly in, or that can be converted to, 2D camera input coordinates. The internal state of the object that is used to describe the objects 2D appearance within the system is composed of the template, the position in 2D camera coordinates and the weighting mask, gained from the indicated area. Eventually, e.g. if the position has been

supplied in 3D coordinates, also the 3D position will be part of the objects state, as well as supplementary information like assumed velocities, sizes, etc.

Once that the tracker has been initialized, the object can be searched in subsequent images by finding those parts of the visual scene where a match between the 2D appearance template and the input features is large. This is accomplished by the 2D tracker, which in our special case is implemented as a state-of-the-art multicue Dynamic Bayesian Tracker which operates incrementally and statistically. The result from this module is a probabilistic map which indicates the probability that the tracked object has currently some particular 2D camera input position and velocity.

If an object's 2D appearance changes considerably over time, it is useful to incorporate cue and template adaptation steps into the tracking procedure, which readjust the 2D appearance template and / or parameters of object specific cues.

As a further step, now the depth is estimated, meaning the 3D distance, between the object and the camera. The central idea of this invention is that two specific (explained in the following) depth estimations are coupled which operate on very different cues and which work in complementary ways to mutually compensate weaknesses of each single estimation. 1) One of them is a direct depth estimation using conventional techniques, like taking advantage of binocular vision. On itself, this direct

depth estimation can be a combination of several different submethods which may have been already integrated with each other. 2) The second technique is the depth change estimation that we can extract from observing a change in size of the 2D appearance of an object in a single camera. The rationale is that an object which is increasing or decreasing its distance to the camera causes a decrease or increase, respectively, of the size of the 2D appearance. If we know the objects depth and the depth velocity, we can estimate the expected size change of an object's 2D appearance. To the contrary, if we know the objects depth and its 2D appearance size change, we can estimate the depth velocity. The two depth estimations (i.e., the direct one and the 2D size change estimation based one) are therefore intertwined and it is reasonable to design a 3D tracking system in such a way that it takes advantage of the combination of these 2 properties. Such a system allows the stable and easy 3D tracking of an arbitrary object without any accurate 3D model of the object itself, relying mostly on its 2D appearance and 2D appearance change.

In particular in combination with binocular techniques, coupling of the two different depth estimation methods provides advantages for a 3D object tracking system. Binocular systems work well in a close distance to the camera depending on the camera baseline length. Depth estimation by observation of the size change of the tracked object, however, can work over a larger range of distances. Both methods can therefore bootstrap each other; as an example, an object whose depth is accurately

measured in the close distance range can be tracked in depth measuring its size change as it moves out of the range of accurate 3D estimation of binocular systems.

The size change of the 2D appearance of an object is extracted in this system using a template matching procedure as described in point 2 of prior art, referred to as 2D transformation estimation. In our case, the system searches for the affine transformations A that provide the best match between the 2D appearance template and the input image. At each timestep t_k, the current transformation state $A(t_k)$ represents the best description of the current 2D appearance of the tracked object in terms of its template, in the sense: "the currently tracked object looks like the 2D appearance template transformed by a certain amount, e.g. rotated by 5 degree and scaled by 10%". The next timestep t_{k+1} transformation state $A(t_{k+1})$ is then estimated by propagating the transformation state $A(t_k)$ into the future and starting a new template matching based transformation estimation search from there. From each affine transformation state, we extract the size λ by approximating it as a composition of a scaling and a rotation operation, in the way

$$A(t_k)=\begin{pmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{pmatrix} \approx \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \bullet \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{pmatrix} \quad .$$

[Eq. 1]

This means that the scale λ can be calculated directly from the determinant of the transformation state matrix. The

results from two consecutive timesteps can then be subtracted to calculate the transformation change, e.g. the scale transformation change $\Delta\lambda$ between two timesteps, which quantifies the size change of the 2D appearance.

The direct depth estimation based on a binocular input was in our case a standard disparity-based dense (i.e., pixel-wise) depth measurement (see point 3 of prior art) followed by an extraction of a single approximate depth for the entire object that is being tracked using the object weighting mask (e.g. by spatial integration of the dense depth measurement with the mask). These measurements were integrated over time again using a Dynamic Bayesian Estimator. An additional, second binocular depth measurement that we incorporated into the system is a direct search of the 2D appearance template extracted from the first camera in the second camera, using e.g. crosscorrelation or the template matching techniques described in point 2 of prior art. From the relative displacements between the positions where the template is found in the left against the right camera, a disparity is extracted that can also be used as basis for a depth measurement.

The direct depth measurement provides some hints on the depth of an object, which is often unreliable. Using the depth from consecutive time steps, we can also extract the velocity in depth of the moving object relative to the observer/the camera platform. If the camera platform itself is moving, and the parameters of this motion are known, then this effect can be compensated, e.g. by

subtracting the camera position and motion from the object parameters to get the object parameters in absolute coordinates.

Additionally to the direct depth measurement, the system delivers a 2D appearance size change measurement, as explained. For approximately rigid objects, and in a calibrated camera system, the relation between its physical size λ_{ph}, its 2D appearance size λ and the depth z is as follows (c is a constant that compresses several system parameters, like the camera focal length, the 2D screen size and the tracking template size):

$$\frac{z}{c} = \frac{\lambda_{ph}}{\lambda} \qquad \text{[Eq. 2]}$$

This expresses the expected fact that the 2D appearance size λ and the depth z are inversely proportional to each other - meaning that for same physical size, objects that are more distant from / near to the camera appear smaller / larger in the camera system.

Figure 2 shows a graphical explanation of the mathematical relationship between the object depth, the object's physical size and its 2D appearance size.

In the system according to the invention, equation 2 couples the direct depth measurement with the depth inference by 2D appearance size change. Internally, we represent a tracked object by a state containing, among

others like velocities, the parameters λ_{ph}, λ and $\{x,y,z\}$. The direct depth measurement delivers, at each time-step, a new estimate for z. The 2D appearance change measurement delivers, at each time-step, a new estimate for λ. The 2D position tracking together with the depth delivers, at each timestep, a new estimate for the global position $\{x,y,z\}$ of the object. The physical size λ_{ph} is an internal parameter of the system that can receive additional hints about its state from other sources, like other sensory measurements, or alternatively prior knowledge about a particular objects physical size, or even supervised input from an interacting person, etc.

The task of the tracking system is now to take the current state parameters $\lambda_{ph}(t_k)$, $\lambda(t_k)$ and $\{x,y,z\}(t_k)$, use them to predict the expected state parameters $\hat{\lambda}_{ph}(t_{k+1})$, $\hat{\lambda}(t_{k+1})$ and $\{\hat{x},\hat{y},\hat{z}\}(t_{k+1})$ for the next time-step based on some dynamical model for the state parameters (like e.g. that the objects physical size is constant, that the object is moving with constant depth velocity and that the 2D appearance size changes according to eq. 2) and couple this with the newly measured estimates for λ (from 2D size estimation) z (from direct depth estimation) and $\{x,y\}$ (from 2D position estimation) to get updated estimations for the new state parameters $\lambda_{ph}(t_{k+1})$, $\lambda(t_{k+1})$ and $\{x,y,z\}(t_{k+1})$, everything under the constraint of equation 2. (In figure 1, the newly measured estimates receive the notation $\bar{\lambda}$, $\bar{z}$ and $\{\bar{x},\bar{y}\}$ to differentiate them from the "true" estimates.)

The direct way to do this is using probabilistic methods, like the prediction-confirmation framework of Dynamic Bayesian Estimators / Recursive Bayesian Filters / Stochastic Dynamic Estimators, for a probability density that comprises at least the the considered state parameters,

$$\rho(\lambda_{ph}, \lambda, x, y, z, ...)$$

improving it over time to get the best estimate of the state parameters, given all past direct depth measurements and 2D appearance size change measurements. In this framework, the newly measured estimates correspond to the probabilistic likelihoods. From the current probability density, we can then extract the most probable parameters λ_{ph}, λ and z that best describe the situation of the tracked object (e.g., by selecting the point of maximal probability, but different methods exist to achieve this). Of course, further object parameters like the object's 2D position and velocity (in camera coordinates) are delivered by the appearance-based 2D tracking system (module 1). Together with the size and depth parameters, the 3D position and velocity of the tracked object can then be determined (module 4). Furthermore, the combination of 3D tracking and keeping track of the position and orientation of the camera (module 5) allows to compensate for egomotion effects, so that objects can be tracked reliably even if the camera system changes its location. In particular, this becomes relevant for applications related to visually guided

robots or cars with mounted cameras for surveillance of other traffic participants, where reliable 3D tracking of objects while the robot / car itself is moving is a necessary capability for building up consistent representations of a visual scene.

REFERENCES

1. Yilmaz, A., Javed, O., Shah, M. "Object tracking: A survey". ACM Comput. Surv. 38(4) (2006) 13
2. Comaniciu, D., Ramesh, V., Meer, P. "Real-time tracking of non-rigid objects using mean-shift". Computer Vision and Pattern Recognition, 02:2142, 2000
3. Comaniciu, V., Meer, P. "Kernel-based object tracking", 2003
4. Lucas, B.D., Kanade, T. „An iterative image registration technique with an application to stereo vision". In *International Joint Conference on Artificial Intelligence* (IJCAI81), pages 674-679, 1981.
5. Tomasi, C., Kanade, T. "Detection and tracking of point features". Technical Report CMU-CS-91-132, Carnegie Mellon University, April 1991.
6. Shi, J., Tomasi, C. "Good features to track". In *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR'94), pages 593-600, Seattle, June 1994.
7. Qian, N. "Binocular disparity and the perception of depth". Neuron 18(3): 359-368, March 1997.
8. Lou, J., Tan, T., Hu, W., Yang, H., Maybank, S.J. "3-D Model-Based Vehicle Tracking", IEEE Transactions on Image Processing, Volume 14, pp. 1561-1569, Oct. 2005.

9. Krahnstoever, N., Sharma, R. "Appearance Mangement and Cue Fusion for 3D Model-Based Tracking", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 2, pp. 249-254, 2003.
10. Giebel, J., Gavrila, D., Schnorr, C. „A Bayesian for multicue 3D object tracking". In *Proceedings of European Conference on Computer Vision*, 2004.
11. Ristic, B. "Beyond the Kalman Filter: Particle Filters for Tracking Applications". Artech House Publishers, February 2004.
12. Arulampalam, S., Maskell, S., Gordon, N.J., Clapp, T. "A Tutorial on Particle Filters for On-line Non-linear/Non-Gaussian Bayesian Tracking". In IEEE Transactions of Signal Processing, Vol. 50(2), pages 174-188, February 2002

3. BRIEF DESCRIPTION OF DRAWINGS

Figure 1 shows an overview of a system for implementing the invention.

Figure 2 shows a graphical explanation of the mathematical relationship between the object depth, the object's physical size and its 2D appearance size.

CLAIMS

1. A method for visually tracking real-world objects by estimating their 3D position and 3D velocity using 2D appearance hints and multicue depth estimations,
the method comprising the steps of:
(1.1.) Taking a camera image measured at time t,
(1.2.) Preprocessing a subregion of the camera image at time t where the object to be tracked is expected using a series of cues to get the input features at time t,
(1.3.) Using the input features at time t to initialize a tracker template by indication of a region of the visual input at time t using external means, with the region gained either from additional algorithms or by user interaction,
(1.4.) Taking the next camera image at time t+dt,
(1.5.) Preprocessing a subregion of the camera image at time t+dt in the same way as in step 1.2,
(1.6.) Using a 2D-tracker on the input features at times t and t+dt to get estimates of the 2D position and 2D velocity of the object's appearance in camera image 2D coordinates, wherein the object is searched in the next camera image at time t+dt by determining a match between the tracker template and the input features,
(1.7.) Using an estimation of the object's depth at time t from an additional cue, measured at two consecutive timesteps t and t+dt in order to approximate the depth change,
(1.8.) Using a 2D transformation estimation on the camera images and/or selected input features at times t and t+dt to extract the relative change of scale / size of the object that is being tracked,
(1.9.) Coupling the approximate estimation of the depth and the depth change of the object from step 1.7 with the

change in scale / size from step 1.8 to improve the depth estimation for the object,
(1.10.) Combining the 2D position and 2D velocity of the tracked object from step 1.6 in camera coordinates with the depth and depth change estimation from step 1.9 and converting it into global 3D coordinates by using camera positioning information, to get the global coordinates of the object that is being tracked,
(1.11.) Using the 3D position to calculate an approximate physical size of the object, and
(1.12.) Iterating the steps 1.4 -1.11 while the object is being tracked, until a stopping criterion is satisfied.

2. The method according to claim 1,
furthermore comprising the step of:
(2.1.) taking into account changes of the position and orientation of the camera in global space to compensate for motion of the camera and / or the platform where it is mounted on.

3. The method according to claims 1 or 2,
(3.1.) where uncertainties are taken into account by using probabilistic methods for the estimation of the object states according to steps 1.6 and 1.7 and / or the coupling according to step 1.9.

4. The method according to any of claims 1-3,
(4.1) where the approximate depth and depth change / depth velocity estimation from step 1.7 occurs incrementally by integrating the results over time.

5. The method according to any of claims 1-4,
(5.1) where instead of a single approximate depth estimation a series of depth estimations based on different cues and / or measurement techniques are used

which are then again coupled with the 2D change in scale / size change estimation as in step 1.9.

6. The method according to claims 1-5,
where the two estimation steps 1.7 and 1.8 influence each other, in the sense that
(6.1.) the estimation of the 2D transformation from step 1.8 occurs by taking into account the expected depth change / depth velocity from step 1.7, meaning that an expected reduced or increased size caused by an increase / decrease in depth is considered in the transformation search procedure,
(6.2.) the estimation of the objects depth from step 1.7 occurs by using prior information about the expected depth derived from the physical size calculated in step 1.11 and the expected change of scale /size of the object being tracked from step 1.8.

7. The method according to claims 1-6,
where
(7.1) the same principles are applied to higher-order derivatives of an objects state parameters.

8. The method according to any of the preceding claims, wherein the result of step 1.12 is forwarded to a visual servoing unit controlling the position and orientation of the tracked object in the input field of the camera means by controlling actuators for adapting the position and orientation of the camera means.

9. A tracking apparatus having camera means supplying a signal to computing means, which computing means are programmed to carry out a method according to any of the preceding claims.

10. A humanoid robot being provided with a tracking apparatus according to claim 8.

11. An automobile, being provided with a tracking apparatus according to claim 9.

12. A computer software program product, implementing a method according to any of claims 1 to 8 when run on a computing device.

The invention relates to a method for the estimation of the dynamic state of a real-world object over time using a camera system, 2D image information and a combination of different measurements of the object's distance from the camera. The 2D image information is used to track an object's 2D position as well as its 2D size and 2D size change using its appearance. In addition, an object's distance from the camera is gained from one or several direct depth measurements. The 2D position and size, and the object's depth are coupled with each other to get an improved estimation of an object's 3D position and 3D velocity, and so get an improved real-world object tracking system, which can be used on a moving platform like a robot or a car with mounted cameras for a dynamic visual scene analysis.

Representative Drawing
Fig. 1

FIGURE 1

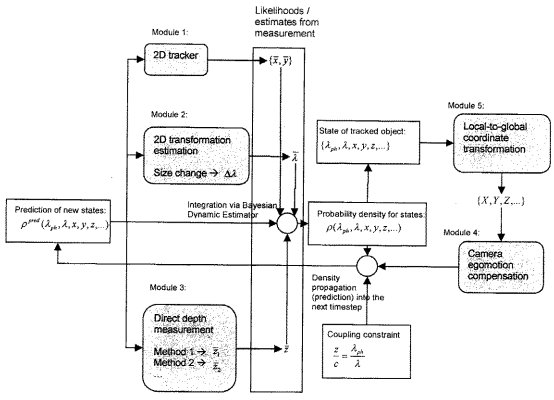


FIGURE 2